Exercise 3.1. The computation of derivatives by finite differences illustrates some of the inevitable tradeoffs in finite-precision calculations.

Consider a twice-continuously differentiable function $f$ of a single variable $x$. The Taylor series expansion of $f$ about the point $\bar{x}$ is

$$f(\bar{x} + h) = f(\bar{x}) + hf'(\bar{x}) + \frac{1}{2}h^2f''(\xi),$$

where $\xi$ lies between $\bar{x}$ and $\bar{x} + h$ (note that $h$ need not be positive!). The simplest form of derivative approximation is the forward-difference formula,

$$\phi(\bar{x}, h) = \frac{f(\bar{x} + h) - f(\bar{x})}{h},$$

where $h$ is called the finite-difference interval. The truncation error $e_T$ in using $\phi(\bar{x}, h)$ (3.2) to approximate $f'(\bar{x})$ is the neglected term in the exact relation (3.1), namely

$$e_T \equiv \phi(\bar{x}, h) - f'(\bar{x}) = \frac{1}{2}hf''(\xi).$$

(a) What does formula (3.3) suggest about how to choose $h$ to make the error in approximating $f'(\bar{x})$ by $\phi(\bar{x}, h)$ as small as possible?

(b) For $f(x) = \sin(x)$, give an upper bound on the absolute value of the truncation error (3.3), expressed in terms of $h$, that is valid for all $\bar{x}$.

(c) Consider the function $f(x) = \sin(x)$ for two values of $\bar{x}$ ($\bar{x} = 0$ and $\bar{x} = 2$) and ten values of $h$ ($h = 10^{-5}, \ldots, 10^{-14}$).

(i) Compute and print $f'(\bar{x})$ to full precision (approximately 16 decimal digits) for $\bar{x} = 0$ and $\bar{x} = 2$. 

In this assignment, please use the two-norm unless otherwise specified.
(ii) For each $\bar{x}$ and each $h$, use ‘format hex’ to print out the IEEE double-precision floating point representations of the computed values of $f(\bar{x} + h)$ and $f(\bar{x})$, and the associated computed difference $f(\bar{x} + h) - f(\bar{x})$.

(iii) For each $\bar{x}$ and each $h$, compute $\phi(\bar{x}, h)$ using formula (3.2), compute the difference $\phi(\bar{x}, h) - f'(\bar{x})$, and print these values.

(d) Explain the results of part (c)(iii), i.e. explain why the computed $\phi(\bar{x}, h)$ does not become an increasingly accurate approximation of $f'(\bar{x})$ as $h$ becomes smaller. What source of error, in addition to truncation error, affects the computed values of $\phi(\bar{x}, h)$? (Note: Your answer should bring in the hex values computed in part (c)(iii).) How might one estimate or bound this second source of error? Does it behave differently at $\bar{x} = 0$ and $\bar{x} = 2$? If so, explain why.

(e) Based on these results, comment on general strategies for choosing $h$ in the forward-difference formula (3.2) to obtain the most accurate computed results.

**Exercise 3.2.** Make up your own examples of $2 \times 2$ real matrices $A$, $B$, $C$, and $D$ with integer elements but with no zero elements such that:

(a) $A^2 = -I$;
(b) $B^2 = 0$;
(c) $CD = -DC$, with $CD \neq 0$.

In each case, please explain how you came up with your examples, and write out each matrix and the associated power (e.g., $A^2$) or product (e.g., $CD$) in full.

**Exercise 3.3.**

(a) If $A$ and $B$ are nonsingular, show that $AB$ is nonsingular.

(b) If $A$ and $B$ are nonsingular, show that $\text{cond}(AB) \leq \text{cond}(A) \text{cond}(B)$, where the condition number of the nonsingular matrix $X$ is given by $\text{cond}(X) = \|X\|\|X^{-1}\|$.

(c) Make up your own examples of nonsingular nondiagonal matrices $A$ and $B$, with $A \neq B$, such that:

(i) $\text{cond}(AB) = \text{cond}(A) \text{cond}(B)$ and
(ii) $\text{cond}(AB) \ll \text{cond}(A) \text{cond}(B)$.

**Exercise 3.4.** In part(b) of this problem, the computed solution $x$ of the linear system $Ax = b$ should be obtained with the Matlab command $x = A \backslash b$ (or any equivalent software that solves linear systems using a backward stable and reliable method and IEEE double-precision arithmetic).
(a) Given a nonsingular matrix $A$ and a vector $b$, let $x$ denote the exact solution of $Ax = b$. If $\|A\| = 1$, $\|b\| = 1$ and $\|x\| = 10^4$, what does this tell you about $\text{cond}(A)$? Why?

(b) Give your own example that consists of the following:

(i) a nondiagonal nonsingular ill-conditioned matrix $A$ for which

$$1 \leq \|A\| \leq 5 \quad \text{and} \quad 10^5 \leq \text{cond}(A) \leq 10^{16};$$

(ii) a right-hand side $\tilde{b}$ with $1 \leq \|\tilde{b}\| \leq 3$ such that $10^3 \leq \|\tilde{x}\| \leq 10^6$, where $\tilde{x}$ is the computed solution of $A\tilde{x} = \tilde{b}$;

(iii) a right-hand side $\bar{b}$ with $1 \leq \|\bar{b}\| \leq 3$ such that $1 \leq \|\bar{x}\| \leq 5$, where $\bar{x}$ is the computed solution of $A\bar{x} = \bar{b}$;

(iv) Explain how you constructed $A$, $\tilde{b}$, and $\bar{b}$.

**Exercise 3.5.** Consider an $n \times n$ upper-triangular matrix $U$ such that

$$u_{11}u_{22}\cdots u_{n-1,n-1} \neq 0 \quad \text{but} \quad u_{nn} = 0.$$ 

(This means that $U$ is singular.)

(a) Assuming that $U$ has the above form, give a general algorithm for computing a nonzero vector $x$ such that $Ux = 0$.

(b) Apply your algorithm to the specific matrix

$$U = \begin{pmatrix}
1 & -1 & -1 \\
2 & 1 & \\
0 & & 
\end{pmatrix}$$

and give the solution you found. What is the general form of all nonzero vectors $x$ satisfying $Ux = 0$ for this particular $U$?

**Exercise 3.6.** Consider the matrix

$$A = \begin{pmatrix}
.932165 & .443126 & .417632 \\
.712345 & .915312 & .887652 \\
.632165 & .514217 & .493909
\end{pmatrix}. $$

(a) What is $\text{cond}(A)$?

(b) Compute the $LU$ decomposition of $A$. What feature of $U$ shows immediately that $A$ is ill-conditioned?
(c) Define $b$ and $c$ as

$$b = \begin{pmatrix} .876132 \\ .815327 \\ .912345 \end{pmatrix}, \quad c = \begin{pmatrix} .876132 \\ .815327 \\ .648206 \end{pmatrix}.$$ 

Compute $\|b\|$ and $\|c\|$. Solve the two systems $Ax = b$ and $Ay = c$, and give the two computed solutions $x$ and $y$ and their norms. Compute the associated residual vectors $r = b - Ax$ and $s = c - Ay$. What difference do you notice in the sizes of $\|x\|$ and $\|y\|$? Is there a comparable difference in size between $\|r\|$ and $\|s\|$? Explain the statement: “$b$ reflects the conditioning of $A$, while $c$ does not”.

(d) Define your own two “random” vectors $d$ and $\tilde{d}$ of unit two-norm. Compute the solutions $z$ of $Az = d$ and $\tilde{z}$ of $A\tilde{z} = \tilde{d}$, and print each solution, its norm, and the norm of the residual. Do your results with $d$ and $\tilde{d}$ reveal the ill-conditioning of $A$? Why or why not? Explain.